Comparing Human Grading to Automated Assessment Technology

Implications for the 21st Century Classroom

Jamey Heit, PhD
Executive Team, Essay Assay, Inc.
Durham, NC, USA
jameyheit@gmail.com

Robin Donaldson, PhD
Executive Team, Essay Assay, Inc.
Paris, France

***Abstract*— We had an experienced professor teach two sections of the same course during the same term. All content covered during the course and all assignments were the same across both courses. Students completed nine written assignments and a final project. The only difference was that in Section A assessment was done by hand as a teacher normally would do. In Section B, all assessment was done using ecree's automated assessment software. Students were not told which section they were in. At the end of the course, we compared student outcomes and evaluations of the professor to determine two things: (a) did one section perform better than the other due to the way they were assessed; and (b) did the students perceive that one form of assessment was better than the other? Our results show that across the board the students in the automated assessment course had better outcomes and showed a clear preference for this mode of assessment.**

Keywords: formatting; automated assessment, student outcomes, technology, writing, instruction

I. INTRODUCTION

To conduct this study, we had an experienced college professor (more than 15 years of experience) teach two sections of the same Introduction to World Religions course. He taught these two sections during the same term. All content covered during the course and all assignments were the same across both courses. Students completed nine written assignments and a final project. The only difference was that in Section A assessment was done by hand as a teacher normally would do. In Section B, all assessment was done using ecree's automated assessment software. Students were not told which course they were in.

There were 16 students in Section A and 15 students in Section B. In the results, Section B will include only scores and analysis from 14 students as one student submitted only a fraction of the required work.

Our purpose in conducting this study was to answer two questions. First, was there a measureable difference in outcomes between the two sections? Second, was there a clear preference for either the human assessment process or automated assessment?

II. TECHNOLOGY OVERVIEW

The technology used was ecree's automated assessment platform, a web-based platform that provides feedback to students in less than a minute. The algorithm that assesses student work is mechanistic, which means that it evaluates student work in the same way that humans do. For example, the algorithm will recognize whether the student includes a thesis statement (statement of purpose, etc.) and then evaluate how good that thesis statement is. The algorithm uses 36 metrics of this kind to score student work and provide feedback.

Using this mechanistic process, the automated assessment technology has the following features: (i) computes accurate percentage grades for essays, (ii) directly measures aspects of quality writing, (iii) provides substantive feedback to the student, and (iv) does not require training. These four characteristics mirror the human process that is used to assess work.

III. STUDENT DATA

	Class A (human assessment average score)	Class B (automated assessment average score)
Assignment 1	89.13	74.86
Assignment 2	88.5	87.14
Assignment 3	91.33	88.14
Assignment 4	88.64	84.86
Assignment 5	91.53	84.36
Assignment 6	91.75	78.07
Assignment 7	86	86.12

Assignment 8	90.63	82.29
Assignment 9	100	85.14
Final Project	90.13	87.36

Table I

	Class A (range of scores)	Class B (range of scores)
Assignment 1	70-100	61-91
Assignment 2	76-100	77-96
Assignment 3	86-100	83-95
Assignment 4	75-100	53-94
Assignment 5	85-100	61-94
Assignment 6	66-100	35-90
Assignment 7	64-96	76-95
Assignment 8	85-98	75-90
Assignment 9	100-100	73-95
Final Project	75-97	82-95
All assignments	64-100	35-96

Table II

Student	Assignment 1 Score	Final Project Score	Difference
1	92	88	-4
2	92	91	-1
3	92	97	+5
4	80	87	+7
5	100	97	-3
6	100	88	-12
7	76	97	+21
8	90	88	-2
9	80	75	-5
10	94	97	+3
11	88	87	-1
12	82	86	+4
13	88	96	+8
14	100	98	-2
15	82	92	+10
16	90	78	-12

Table III

Student	Assignment 1 Score	Final Project	Difference

		Score	
1	83	88	+5
2	85	85	0
3	85	93	+8
4	91	88	-3
5	70	88	+18
6	70	81	+11
7	89	95	+6
8	87	86	-1
9	76	82	+6
10	81	86	+5
11	61	88	+21
12	82	87	+5
13	88	88	0
14	81[1]	88	+7

Table IV

IV. Analysis

As Table I shows, there was a clear difference in the scores for each assignment based on which section students were in. Figure I shows each section's class average score for each the ten total assignments.

There is a clear trend of higher scores for students in Section A. The data are unambiguous that human grades will result in a higher overall grade for the course. The only assignment where the average score in Section B was higher is Assignment 7. Unlike the other assignments where students in Section A have a clearly higher average score, students in Section B realized only a 0.12 point "advantage" over their peers in Section A.

Assignment 9 shows clearly that human assessment can result in higher scores for students. This assignment was a reflection on concepts learned in the course. Students in Section B receive scores that were in line with how their work was assessed throughout the course (as one would expect with the consistency in assessment an algorithm provides). Students in Section A, however, were clearly given automatic scores. Moreover, those automatic scores were perfect. The clear suggestion is that the human completed the final grading in the course in a way that was meant to benefit the students without basis in the actual work completed.

One point that stands out is that the highest score for the vast majority of assignments is a 100. In fact, 20% (30 out of 150 total papers submitted) of all grades given were a perfect 100[2]. It would be highly unlikely for 20% of all papers to have no room for improvement, which shows that the human graded was not evaluating student work closely. A more likely

expectation would be *no* perfect scores, as is the case in Section B.

Just as there is a tendency to score too high, the human section of the course also assigned *only* passing grades. The lowest score is a D. Further, only 1.3% of all grades (2 out of 150) were given a D score. There is not a single failing paper in Section A.

The conclusion is that human scoring not only allows for higher scores at the top of the grading sale, but also that the human grader makes sure that grades aren't too low.

A final point about the scores in the two sections should be mentioned. The difference between the first assignment's score and the final project score offer insight into how much students improved over the term. A student who has a lower initial score but a higher final project score is likely showing improvement over the course.

Table III shows each student's Assignment 1 score alongside her/his final project score. Similarly, Table IV shows each student's Assignment 1 score and final project score. For both tables, we have included in the final column the difference between initial score and final project to highlight the overall improvement for each student.

A look at the range of scores for each assignment strengthens the claim that students in Section A had higher scores because they had a human assessing their work. Table II shows the range of scores for each assignment.
The data suggest that students in Section A were not improving as much based on the bump in scores they received from the human grader. 56% of students did worse on their final project than on their initial paper. The average decline for these students was 4.7 points. Of the seven students who showed improvement, the average increase was 8.3 points. When students did improve over the course of the term, the improvement was significant. However, the majority of students did not exhibit this improvement.

In Section B, however, there is a significantly different trend. Only 14% of the students showed a decrease in score between their first assignment and their final project. The average decrease for these students was only 2 points. 71% of students improved over the course of the term with an average improvement of 9 points.

The key takeaway regarding the benefit of automated assessment on outcomes, then, is that it helps more significantly students improve. In both sections, student improvement was roughly the same in terms of change in score. The significance is in the number of students who showed improvement. For those assessed by the human, less than half of the class improved. For those assessed by an algorithm, more than 70% improved.

V. Student Evaluations

Just as there was a clear trend in student outcomes between Sections A and B, there was also a noticeable pattern in how students evaluated the teacher. We consider three pedagogical metrics and three metrics that speak to the student experience with assessment. In all cases students were asked to rate the professor on a scale of 1-5 (1 being the lowest).

The pedagogical categories include: Overall Satisfaction and Instructor Was Knowledgeable.

In Section A, 71% of students rated the class as a 5, 14% rated the class as a 4, and 14% rated the class as a 1. The low ranking was consistent across all categories, so we conclude that this is a single student who simply gave the professor the lowest possible marks. In our analysis, then, we exclude this student's ranking as an outlier not based on how the professor conducted the course.

In Section B, 73% of students rated the class as a 5, 18% rated the class as a 4, and 9% rated the class as a 3.

Based on the overall rankings, the professor seemingly offered a consistent classroom experience for both sections. If there were perceptible differences in how the professor presented material and managed discussions, we would have expected to see such differences manifest in the overall rankings. The similarity between the two sets of evaluations suggests that the quality of the course was consistent.

The same pattern emerges in looking at the results for Instructor Was Knowledgeable. For Section A, 73% of students rated the professor a 5 and 14% rated the professor as a 4. In Section B, 73% rated the professor as a 5, 18% rated the professor as a 4, and 9% rated the professor as a three. As with the Overall Satisfaction metric, these results show that the professor's command of the material was perceived more or less the same across the two sections. Based on these evaluations of the professor's pedagogy, the conclusion is that there was no significant difference in how the professor taught the course.

The three metrics to consider are: Feedback Was Specific to the Assignment, Additional Feedback Was Given, and Feedback Helped Me Improve.

For Section A, students conveyed that their experience with assessment was not as good as the overall course. Only 50% of students rated the professor a 5 for Feedback Was Specific to the Assignment. 25% of students rated the professor a 4, 12.5% rated the professor a 3, and 12.5% rated the professor a 1.

The results for the other two metrics regarding assessment showed a similarly pattern. For both Additional Feedback Was Given and Feedback Helped Me Improve, 62% of students rated the professor a 5. 12% of students rated the professor a

4, 12% rated the professor a 3, and 12% rated the professor a 1.

When considering the metrics that relate specifically to assessment where there was a difference, the numbers are telling. Given the bias towards higher scores overall in Section A, we expected to see a more positive response to assessment from the human. Instead, there is clear evidence that students were less satisfied with the assessment in Section A than with the overall course.

In Section B, students responded much differently. For Feedback Was Specific to the Assignment, 69% of students rated the professor a 5, 23% rated the professor a 4, and 8% rated the professor a 3. There was, then, a much better perception of the professor's assessment in that a higher percentage of students rated the assessment as superlative and no students were clearly dissatisfied.

The more significant difference between Sections A and B occurred in the last two assessment metrics. For Additional Feedback Was Given, students in Section B clearly felt they had a better experience. 84% of students rated the professor a 5, 8% rated the professor a 4, and 8% rated the professor a 3. One clear takeaway is that the algorithm provided students with more feedback than they were expecting.

Likewise, there was a noticeably higher evaluation of the impact the feedback had on student learning. For Section B, the Feedback Helped Me Improve results showed that 79% of students rated the professor a 5 and 21% rated the professor a 4. There were no scores lower than a 4. The algorithm provided not only a more substantive set of feedback to students, it was also perceived as having a much better impact on the students' perspective on their learning.

VI. Student Comments

At the end of the professor evaluation, students had the option to write-in additional comments. These comments mirror the trends noted in the previous section. On balance students in both sections enjoyed the course, but there was a clear perception of better assessment in Section B.

A look at the positive responses for Section A are general in their praise of the course. The four comments are: "His instructions are clear and direct, his comments clarify and educate"; "I was not sure what to expect with this course. I feel that the instructor helped me to look at things in a different view. I enjoyed the course"; "I can confidently say that my instructor did his best. I am very satisfied with his work"; and "The course is a great one with lots of new experiences for me; and i think i will succeed in the long time."

Only one student mentioned a specific attribute of their positive experience. The instructions that are noted could be relevant to assessment, but any such link would be speculative based on the student's language.

Importantly, the negative comments about the professor in Section A do highlight specific concerns: "He would grade a 46 out of 50, or a 48, or something close to 50 but rarely 50. It seemed as if he were compelled to skim every grade just a little"; and "I feel as though we did not have enough time to really learn it to many application and reflection assignments each week. And like I said the instruction was not clear." We assume the latter comment here is from the outlying student mentioned in the previous section and exclude it from our analysis here. The relevant feedback is one of perceived biased in the assessment process. The student identifies an inconsistency, but it has to do with not getting perfect scores. Given the percentage of perfect scores given in the course, this comment suggests that the scoring bump students in Section A received created similarly inflated expectations about what grades *should* be received.

A look at the positive comments from Section B once again shows a different student experience. The positive comments are: "His feedback on each paper that needed correction or improvement was specific and helpful. I used his feedback to work on weaknesses in my writing and how to look differently at the resources provided"; "My prior classes were very slow to grade and even at my local community college I waited until the last week of class to know my grades from the 2nd week. What a fantastic experience;" "My Instructor was awesome!!!!"; "This is the first class I've taken where I felt like the instructor took the time to provide me with specific areas of improvement"; and "I know that with all of this information and your feedback (to get us to think outside of the box) will help overall understand the many different cultures and linguistics I will encounter on a daily basis. Again, thank you for all the great feedback and responses it has been a great course!"

There is a clear theme of better feedback in these comments. The students recognized in their experience with automated assessment each of the benefits the technology offers: more substantive feedback, faster turnaround, and a direct effect of assessment on learning. These comments suggest very clearly that the generally more positive student experience towards automated assessment is based on specific enhancements that the human did not provide.

The negative write-in comments for Section B do show a pattern. The comments are: "One thing I think I would have liked is an actual rubric that we could see"; and "I would of liked to seen more tailored rubrics for the individual assignments. There was just one general rubric for all the assignments." The specificity of concern about the rubric is a curious commonality given the generally positive experience students in Section B had with automated assessment. We

posit that this interest in a more defined rubric stems from the students' having better feedback on their papers. Because they were given more insights into how their work could be improved, they wanted a more detailed breakdown of what constituted improvement.

VII. Key Takeaways

After the two sections had ended, one student from Section B sent an email to the professor to offer unofficial feedback. The student emphasized the positive experience with the assessment process. The email said: "I just wanted to thank you for the great feedback you've been giving on my assignments. In my past courses, most feedback received, if any at all, were simple "Good Job!" statements. Thanks!"

These final words capture the key takeaways from our study. Holding everything else constant, human assessment was perceived as less impactful than automated assessment. This trend is born out in the grades for the course assignments. A general conclusion, then, is that automated assessment offers significant pedagogical value to students in the classroom Setting.

VIII. Conclusion

To conclude this study, we offer a couple of final comments. The first is to recognize a way to critique the results here. We accept that claims of bias are potentially valid critiques of these results. Though the professor is an established teaching professional, we accept that subconsciously his decisions could have been biased. We see no evidence of this in the data, but we allow that it could have been present.

Given the ambiguity possible bias introduces into these results, we offer two further comments. The first is to emphasize that we consider these results preliminary. The conclusions we have reached are promising and we hope they will generate meaningful discussions about the role of automated assessment as a learning resource.

The second comment is to explain the steps we are taking to test these matters further. To see if these results are repeatable and to remove the question about bias on the professor's part, we have reached an agreement with a major research university to conduct a significant third-party study of this technology. In an upcoming term this university will conduct a similar experiment across multiple sections of the same course with 300+ students. We are confident we will see similar outcomes and we look forward to generating more data about the pedagogical value of automated assessment technology.

IX. Notes

[1] All raw data is available to those interested. Please email the author(s) if you would like to see the data.

[2] Student 14 did not turn in Assignment 1, so we have used Assignment 2 as the initial score.